

DOCUMENT RESUME

ED 172 806

IR 007 441

AUTHOR Ehrlich, Lisa R.
TITLE The Design, Development, and Evaluation of an Evaluative Computer Simulation.
PUB DATE Mar 79
NOTE 21p.; Paper presented at the Annual Convention of the Association for Educational Communications and Technology (New Orleans, Louisiana, March 1979); For related document, see IR 007 416 .
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Clinical Diagnosis; Computer Oriented Programs; *Evaluation Criteria; Evaluation Methods; Evaluative Thinking; *Medical Care Evaluation; Medical Education; Medical Treatment; *Reliability; *Simulation; *Validity

ABSTRACT

This paper discusses evaluation design considerations for a computer based evaluation simulation developed at the University of Iowa College of Medicine in Cardiology to assess the diagnostic skills of primary care physicians and medical students. The simulation developed allows for the assessment of diagnostic skills of physicians in the selection of medical history taking, physical examination, and the ordering of diagnostic laboratory tests. Evaluation criteria used in assessing this simulation are discussed, scoring considerations are explained, and the issues of evaluation validity and reliability are explored. (RAO)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED172806

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

The Design, Development and Evaluation of an Evaluative
Computer Simulation

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Michael Simonson

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM."

Presented at the Association for Educational Communication
and Technology National Convention in New Orleans 1979

Research and Theory Division

by

Lisa R. Ehrlich, M. A.

University of Iowa College of Medicine

IR007441

The Design, Development and Evaluation of an Evaluative
Computer Simulation

Lisa R. Ehrlich, M. A.

University of Iowa College of Medicine

This presentation will first compare and contrast the differing purposes of instructional and evaluative computer based simulations. It will go on to discuss the development of a cardiology simulation that will be used as a referent for discussion. The design and structure of this referent simulation will then be presented. The major portion of this paper will discuss the evaluation design of this cardiology simulation and some measurement issues of computer based evaluative simulation.

The technique of computer based simulation has been successfully utilized in an instructional mode for some time in many academic disciplines. Within the past two decades, much preliminary work has been conducted to investigate the use of simulation techniques, on and off line, as evaluative instruments. While the technique of constructing a model of reality for simulation purposes remains essentially the same for either an evaluative or an instructional simulation, the instructional designs differ vastly. This paper will first discuss the differences in instructional design necessitated by the differing educational purposes of these two types of computer based simulations. Once these differences are established, then development and evaluation issues will be presented and discussed using an ongoing research project at the University of Iowa College of Medicine as a model reference. The final portion of this paper will discuss considerations relevant to scoring of evaluative simulations and the issues of validity and reliability.

Purposes of Instructional and Evaluative Simulations

The purpose of an instructional computer based simulation is to allow the student to interact with a simulated model and to learn how this simulated environment reacts to differing conditions imposed on it. Most instructional simulations have as part of their design some instructional segments, or, programs that help to guide and teach the learner various principles or conditions present in this simulated environment. Many of these instructional segments are tutorial in that they teach the student some of the salient features of the model, or simulation,

and allow the learner to progress from very basic principles, or concepts, to more complex aspects of the system of interest. The more sophisticated instructional simulations utilize remedial branching as part of their design to reinforce learning. Some of these complex instructional simulations have internal test items that are used as a check for learner comprehension and understanding of the concepts presented as part of the tutorial package.

Once the learner has successfully grasped the basic principles and concepts, s/he is first taught how to interact with the model, or simulated environment, and ultimately allowed to experiment and manipulate this model. In this phase of the instructional simulation the learner sets up the varying conditions and constraints and then explores how the model reacts given these determinants.

To summarize, the general design of instructional computer based simulations is to first create a model of the object system, to develop a series of instructional tutorial programs which teach various concepts and principles of the object system and to ultimately allow the learner to interact with this model by manipulating and varying the conditions. The primary purpose of this type of computer based simulation is instructional.

By contrast, the primary purpose of an evaluative simulation is to test, or measure knowledge, skills, and/or competencies within a specific discipline or subject area. Its purpose is not instructional; any learning that does occur is secondary to the main purpose of evaluation. The same principle holds here as

with any written cognitive test instrument: whatever a student learns while being tested is purely secondary to the purpose of evaluating what a student "knows" or "understands".

Development of a Cardiology Evaluative Simulation

As stated earlier, the evaluative simulation which will be used as a reference is one developed at the University of Iowa College of Medicine in Cardiology. The purpose of this computer based evaluative simulation is to assess both diagnostic and therapeutic skills of primary care physicians and medical students on acute myocardial infarction. The design of the simulation allows for assessment of diagnostic skills of physicians in sections of medical history taking, physical examination and in the ordering of diagnostic laboratory tests. Patient management skills are evaluated for patients who have uncomplicated illness - in this case myocardial infarction, as well as in section dealing with common complications of the disease: ventricular arrhythmias, bradycardia, hypertension, cardiogenic shock, and cardiac arrest and resuscitation.

In June of 1977 a development team consisting of a content expert (cardiologist), a medical liason (physician's assistant and/or a senior medical student), and an instructional designer-programmer began a two year project to develop an on line evaluative simulation in cardiology. A review of the literature was first done to ascertain the state of the art. Once this was accomplished, the content expert and the instructional designer met to begin the content development of this simulation. These

"brain storming" sessions lead to an overall design of the "simulated patient", or environment, which was then initially flowcharted. At this point a senior medical student was brought in and introduced to the concept and technique of evaluative simulations and flowcharting. For the first six months, many meetings occurred between the three members of the development team involving discussion of the case and the ramifications of diagnosis and treatment. The medical men had no computer or instructional design background and the instructional designer had no medical background, so, each member of the team learned what exact information was necessary for the development of each part of the simulation.

Since the major portion of this presentation is to center around measurement considerations of evaluative simulations, the design and development aspects will not be detailed. Suffice it to say that the design, development, programming, and the debugging processes of the simulation took one and a half years to accomplish. The programming was done in BASIC on a Hewlett-Packard 2000 Access minicomputer located at the University of Iowa Computer Assisted Instruction Laboratory.

Design and Structure of the Cardiology Simulation

Prior to consideration of some measurement issues, a short discussion of doctor (user) interaction with the simulation model is warranted. Since the purpose of this instrument is evaluative, the user is essentially allowed to progress through the simulation at her/his own rate and with minimal structure

imposed. The program was designed to allow the doctor to follow whichever path(s) s/he chooses and is consequently forced to deal with whatever complications arise. Therefore, there is no direction or guidance imposed.

There is no instructional remediation incorporated within this simulation, since instruction is not the intent of this instrument. The feedback the doctor receives is twofold: one form is the result(s) of treatment(s) that were ordered and how this treatment(s) affected the patients' condition. The purpose of this feedback is primarily informational rather than instructional. The second form of feedback is both informational and instructional. This feedback occurs either at the end of the simulation when the patient is successfully treated and released from the hospital, or, if during the simulation the doctor, through incorrect diagnosis or treatment kills the "patient". The program records which sections the doctor has completed and once the simulation is terminated, successfully or unsuccessfully, gives detailed feedback in terms of diagnosis, recommended treatment and rationale. This feedback is included as a summary of the case.

The structure of the simulation is as follows: the simulation is broken down into an introductory office scenario, and a series of management programs on premature ventricular contractions (PVC), bradycardia (day 2), hypertension (day 3), and a final program on cardiogenic shock (day 4). There are also two different cardio pulmonary resuscitation (CPR) programs; one for management of cardiac arrest in the coronary care unit and the

other for treatment of cardiac arrest in an office or unmonitored ward. There is continuity of presentation of these different management programs. The doctor will proceed through the simulation without realizing that these programs constitute different sections.

The actual logistics of this sequencing will not be addressed in detail as it is complex and not the primary intent of this paper. It is important to note that at many points throughout the simulation the patient can die due to improper management. Consequently, some of the doctors will not get to all the sections of this simulation if the patient dies prematurely.

Validity Issues of Computer Based Evaluative Simulation

With the purpose and structure of this simulation in mind, attention is now turned to various measurement considerations and the research design that will be used in evaluating this simulation. It should be noted that the measurement issues that will be raised are not specific to this particular simulation, but, generalize to most evaluative simulations on and off line.

The first step in the evaluation process is validation. Therefore, the primary purpose of the research design is to validate the simulation. The audiences that will be used are not themselves being evaluated. It is their performances on the simulation and on various external validation instruments that are of utmost concern in ascertaining the validity and the reliability of this simulation.

There are four different audiences that will be used to validate this evaluative simulation. These audiences consist of a group of senior medical students at the University of Iowa, a group of primary care physicians who are returning to the University of Iowa for a five day workshop as part of a continuing medical education program, the staff cardiologists at the University, and 20 to 30 internists located throughout the state of Iowa. A general overview of the validating instruments will be discussed as well as the research designs for each group. The data handling and statistical analysis will not be presented since they are technical rather than conceptual issues.

There will be three external validating instruments used: 1) a cognitive written multiple choice exam 2) a subjective clinical rating scale and 3) a medical audit.

The cognitive written multiple choice exam will be used with the senior medical student group and the group of primary care physicians. This exam will contain questions which address the doctors' knowledge level of cardiology and cardiovascular disease. There will also be a series of questions used which address more skill related areas of cardiology. These questions will force the doctor to go beyond the knowledge level, to apply and analyze specific circumstances in order to diagnose and manage patients in various situations.

This exam will be tailored to correspond to the major management points addressed in the simulation. By constructing the exam in this manner, it will enable direct comparison of specific performances on the simulation and on related test

items. It is felt that this is the only means of using a cognitive exam as a validating instrument. To help validate the simulation, the scores obtained on the simulation will be compared with the exam results to assess 1) if there is a relationship between scores on each part of the exam and performance on the corresponding segments of the simulation and 2) if the results on the total simulation breakdown into the same relative rankings as they did on the exam.

The subjective clinical rating scale will only be used with the medical student group. The major problems in using this subjective scale as a validating instrument are inter and intra rater reliabilities, as well as the actual observation time each rater has with each student. There is unfortunately no way to regulate the amount of time a rater has with each student, as most of these encounters are spontaneous. The consistency with which a rater rates individual students is a problem with any observational instrument and an attempt will be made to control for this. This instrument will be constructed in a manner to enable the students to be rated on specific diagnosis and management criterion that corresponds to major diagnostic and management points in the simulation.

The ultimate criterion for validating the simulation is to assess how well actions and decisions taken on the simulation match actions taken in real life. There exists in medicine a means by which this assessment can be accomplished, that is by conducting a medical audit. A patient's chart contains all the relevant information on the diagnosis and management of a case.

It is with this record of information that a case can be reconstructed and the major medical decisions determined. A medical audit is the means by which a case is reconstructed.

Twenty to thirty internists throughout the state of Iowa will be participating in this phase of the validation. With their permission an audit will be conducted on those patients' charts that directly relate to the content of the simulation. Specific comparisons will be made between performance in real life, as assessed through the audit, and the diagnosis and management of the simulated patient.

The criterion related, or more specifically, the concurrent validity of this simulation will be assessed by using the cognitive written test instrument. The relative rankings on the exam will be compared with the relative rankings for both groups on the simulation.

It is assumed that this written cognitive instrument is both valid and reliable and is in fact measuring the same skills that the simulation is measuring. This is a critical assumption. A simulation does not duplicate reality, it only imitates it. Because of this, in assessing the concurrent validity of this simulation it is only possible to determine some aspects of physician performance. This, or any simulation can only give a measure of how a physician is capable of behaving not how well s/he will behave in the actual situation. It is important to bear this in mind when determining the concurrent validity of simulations.

The clinical ratings of the student group will also be used to assess concurrent validity. This too will use relative rankings of the students on individual sections of the clinical rating scale. The results on specific sections of the simulation will be compared with the ratings. It is hoped that there will prove to be a relationship between how the students perform in the clinic and how they perform on the simulation.

Finally, the results of the medical audit will be compared with the physician results on the simulation. It is hoped that this will be the ultimate validating criterion. If it is found that the physicians' performance on the simulation relates to their performance in practice, then some proof of concurrent

validity can be established.

It seems a logical construct to assume that a physician who has both diagnosed and managed patients in both clinical and office situations will perform "better" on the simulation than medical students who have had limited experience. It also seems logical to assume that cardiologists should perform "better" on the simulation than either of the above groups.

Construct validity will be determined by comparing the performance of these differing groups. Construct validity will be tested as follows: 1) the primary care physicians should perform "better" as a group on the simulation than the group of medical students and 2) the Cardiologists should perform "better" as a group on the simulation than either the physician or student group. If the simulation is able to differentiate the performance of these differing groups, then it may be said with some certainty that the simulation possesses construct validity.

It is assumed that there is a significant difference between the groups tested. This will be determined by comparing the performance of the student and primary care physician groups on the written cognitive instrument. It is assumed that the group of cardiologists is significantly different than the above two groups. If significance is not found, then validating the simulation by this means is not appropriate.

The content validity of this simulation will be determined by an expert team of cardiologists.

Scoring of Computer Based Evaluative Simulations

One of the most crucial and fundamental problems of any

evaluative simulation is scoring. All the external validating instruments that were discussed rely on a reliable scoring mechanism. No matter how well these instruments are constructed, and are psychometrically sound, if the scoring of the simulation is not both reliable and realistic the validity can not be determined. Nor, does it matter how much face validity the simulation possesses if criterion related validity can not be shown. Critical to proving validity of this simulation is the development of a sound scoring algorithm.

It is crucial that the scoring algorithm realistically represent the clinical ramifications of treatment. The purpose of scoring the simulation at each data option point is to obtain an overall numeric score on the simulation for each doctor and also obtain a series of subscores on each section of the program.

This simulation contains a multitude of major data option points all of which will never be encountered in any one patient. A major data option is defined as a treatment list which is encountered when the doctor opts for medical or surgical intervention, or, as a management list which usually contains a list of options allowing the doctor to perform a physical, take a history, order laboratory tests, and/or begin medical intervention.

The scoring will be accomplished in two independent phases. The first phase will be to have a number of cardiologists take the simulation. The individual physician paths will be monitored and placed in a file. These files will consist of individual physician paths and a cumulative frequency record of all the data options that were chosen.

Once these physicians have gone through the simulation, their paths will be reconstructed on paper in the form of decision tables. Since there is no one optimal path through the simulation, the cumulative paths determined by these cardiologists results will serve as the standard "optimal" paths. Once these standards are established, a weighting scale can be instituted. It is important to note that for any major data option there exist a multiple number of routes that could be taken to reach that option point. In other words, it is possible to reach a data option by one route and choose drug 'A' for treatment and receive the maximum weighted score. It is also possible to reach that same data option and choosing the exact same drug, by taking a different and less desirable route and not receive the maximum weighted score. This indicates the need to standardize optimal pathways first.

There are hundreds of treatment and management lists incorporated within this simulation. It would be almost impossible for a group of experts to weight (score) every option within every data list. It is because of this reality that the decision was made to limit and identify only the major data option points throughout the simulation. These data options constitute critically important points for diagnosis, treatment and mangement. It is these major data options that will be weighted.

A decision table will be constructed for each major data option point and all possible routes that could have been taken to reach that particular option will be specified. A 10 point weighting scale will be used to judge the plausibility of each

data majora option within the simulation. A team of experts will be employed to rate EACH major data option on a +5 to -5 scale [+5 critical (essential to do) to -5 very harmful (essential to avoid)]. From the predetermined standard the +5 weighted score will be cited for each specific option on the optimal path. It will be the job of this team to weigh the other options for each major data option list. It is quite possible and probable for any one option within a data option list to have differing weights due to different paths.

While the scoring is considerably complex, it is essential that it be done. The team of experts will work independently rating each option of each major data option list; the results of the ratings will be compiled, and the final ratings for each option will be determined. If the ratings are not consistent on any one item, the highest and lowest ratings will be discarded; and the remaining scores will be averaged to determine a final weight for the option.

Reliability of Computer Based Simulations

The last measurement consideration is that of assessing the reliability of this simulation. The ideal way of measuring reliability is to construct and develop two parallel evaluative simulations and look at individual performance on these parallel forms. Not only is it extremely difficult to develop two parallel simulations measuring the same knowledge and skill levels, but, this form of reliability estimation falls victim to all the intervening and confounding variables of any parallel forms situation. This procedure is also extremely costly and

time consuming, considering that it took one and a half years to develop this present simulation.

Another, more viable procedure for establishing reliability is using a test-retest design. While this is certainly possible, this procedure is susceptible to all the intervening and confounding variables that this estimation presents.

Internal reliability estimates such as Crobach's Coefficient Alpha or KR20 all make the fundamental assumption that the individual test items are discrete and identifiable. That is, a 50 item exam has 50 discrete, individual test items. This is not the case in an evaluative simulation. There are no test items that can be individually identified.

McGuire and Babbot (1967) in their article entitled "Simulation techniques in the measurement of problem solving skills" (1) have looked at this issue of reliability. They state that there are four characteristics that have the effect of reducing reliability as it is commonly estimated: 1) the data options are differentially weighted 2) data options are interdependent 3) there are differential amounts of feedback obtained by the examinees which results in dissimilarity among the students with respect to the nature of the problem posed by any given data option and 4) an individual student can be denied the opportunity of responding to many of the data options because of the particular decisions that s/he opts to make (i. e., may follow many different alternative paths).

It is because of these four factors that reliability estimation is extremely difficult. McGuire and Babbot take the stance that reliability of a measure can be viewed as an attribute

related to it's generalizability with respect to different universes. Of interest are the estimates of the extent to which a particular set of scores on the simulation is generalizable to many possible similar tests. Here they are referring to the characteristic that Cattell has included in his concept of "consistency of measurement" and which he specifically refers to as "homogeneity". They recommend using any of the three following techniques which have yielded essentially similar results from their own studies; Angoff formula 12, Cronbach's Coefficient Alpha, and principal component analysis. All will estimate the consistency of measurement or generalizability of results across tests.

While it is possible to obtain a reliability estimate according to McGuire and Babbot, the four characteristics inherent in a simulation that reduce reliability have yet to be resolved. It seems possible that a coefficient can be derived which is sensitive to the multitude of confounding variables that exist in this technique. What is needed is more theoretical research in this area by those familiar with this technique and the areas of reliability estimation.

Summary

In summation, evaluative computer based simulations are both exciting and challanging measurement instruments for any eduaction environment. This mode of evaluation offers an innovative means of assessing a learners' comprehension, knowledge, and skill levels within any subject area. While there are many unresolved measurement issues, the foundation is laid

for researchers to explore these vital area of interest.

Footnotes

- (1) McGuire, C.H., & Babbot, D. Simulation techniques in the measurement of problem solving skills. Journal of Educational Measurement, 1967, 4, 1-10.